Lectures on Mathematics in the Life Sciences
Volume 17, 1986

Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences

Simon Tavaré

ABSTRACT. This paper concentrates on statistical aspects of the estimation of substitution rates and divergence times on the basis of DNA sequence data. A new method of estimation is suggested, and exhibited using data from serum albumin and α-fetoprotein. The divergence time of rat and mouse is estimated using a tree calibrated by the human-rat divergence time. Some inherent difficulties in these methods are highlighted by statistical analysis of the sequences.

I. INTRODUCTION

This paper is concerned with probabilistic and statistical questions relating to the estimation of substitution rates and divergence times on the basis on DNA sequence data. Suppose that we have two functionally homologous genes taken one from each of two species. On the basis of just the observed differences in base composition of the two sequences, we want to estimate the time of divergence of the two species, and to estimate parameters of the evolutionary process that led to these observed differences.

A general model of this process of mutation should take account of the relative roles of substitution, insertion and deletion, duplication and transposition as forces that change the structure of genes over time. Here I will focus only on the effects of substitution, the replacement of one base by another. I will also

0075-8523/86 $1.00 + $.25 per page

assume that the two sequences under comparison are of the same base length n, say.

We will label the four bases A, C, G, T by 1, 2, 3, 4 respectively, and let $(X_i(t), Y_i(t))$ denote the bases that occur at the i^{th} position $(1 \le i \le n)$ in species 1 and 2 respectively, t time units after divergence. The assumption of divergence from a common ancestor means that

$$X_i(0) = Y_i(0), \quad i = 1, 2, \ldots, n. \tag{1.1}$$

This article is divided into six sections. In Section 2, we review some stochastic models for the behavior of the process $\{(X(t), Y(t)), t \ge 0\}$ which describes the base composition of two homologous nucleotide positions in the sequences. Section 3 and 4 discuss some statistical questions concerning estimation of evolutionary parameters and goodness-of-fit tests for these models. The methods are illustrated with respect to the serum albumin and α-fetoprotein genes of man, mouse and rat. In Section 5 we treat the problem of estimating the divergence time of two species on the basis of sequence data when the phylogeny length is calibrated by a more distant sequence of known divergence. Section 6 contains some concluding comments.

II. STOCHASTIC MODELS OF SUBSTITUTIONS

We need to model the stochastic behavior of the process $\{(X(t), Y(t)), t \ge 0\}$ in which t denotes time of divergence from the common ancestor, and $X(\cdot)$, $Y(\cdot)$ denote respectively the nucleotide in homologous positions in sequence 1 and sequence 2. As in (1.1), we have $X(0) = Y(0)$ and subsequently $X(\cdot)$ and $Y(\cdot)$ evolve independently.

The substitution process $\{X(t), t \ge 0\}$ can be described by the transition functions

$$p_{ij}^X(t) = P[X(t) = j | X(0) = i], \tag{2.1}$$

with a corresponding function for the $Y(\cdot)$ process. If we define

$$f_{ij}(t) = P[X(t) = i, Y(t) = j | X(0) = Y(0)], \tag{2.2}$$

then our assumptions readily give

$$f_{ij}(t) = \sum_{\ell=1}^{4} \pi_\ell p_{\ell i}^X(t)\, p_{\ell j}^Y(t) \tag{2.3}$$

where

$$\pi_\ell = P[X(0) = \ell] \equiv P[Y(0) = \ell]. \tag{2.4}$$

It is often assumed that

$$p_{ij}^X(t) = p_{ij}^Y(t) \equiv p_{ij}(t). \tag{2.5}$$

Under this assumption, if we write $F_t = (f_{ij}(t))$, and $P_t = (p_{ij}(t))$, then (2.3) becomes in matrix notation:

$$F_t = P_t^T F_0 P_t, \quad t \ge 0 \tag{2.6}$$

where $F_0 = \text{diag}\{\pi_1, \pi_2, \pi_3, \pi_4\}$.

It remains, of course, to specify P_t. The model most frequently used is the case in which $\{X(t), t \ge 0\}$ is a continuous-time time-homogeneous Markov chain, in which case we have (cf. Karlin and Taylor (1975), Ch. 4)

$$P_t = e^{Qt} := \sum_{n=0}^{\infty} Q^n \frac{t^n}{n!}, \quad t > 0. \tag{2.7}$$

Here $Q = (q_{ij})$ is the generator of $\{P_t\}$; Q satisfies

$$q_{ij} \ge 0 \ (i \ne j); \ q_i = -q_{ii} \ge 0; \ Q\underline{1} = \underline{0}, \tag{2.8}$$

where $\underline{1}=(1,1,\ldots,1)^T$, $\underline{0}=(0,0,\ldots,0)^T$. We also make a stationarity requirement by assuming that $\underline{\pi} = (\pi_1,\ldots,\pi_4)$ satisfies

$$\underline{\pi}\, Q = \underline{0}\ . \tag{2.9}$$

If we also assume that $\{Y(t),\ t \geq 0\}$ has the same stochastic structure as $X(\cdot)$ (so that, in particular, (2.5) holds) then the marginal distributions of $X(t)$ and $Y(t)$ are identical (and equal to $\underline{\pi}$) for all t.

From now until the end of Section 3, we will assume that $X(\cdot)$ and $Y(\cdot)$ are stochastically identical. The evolutionary parameter of interest is then the compound parameter K defined by

$$K := 2t \sum_{\ell=1}^{4} \pi_\ell q_\ell . \tag{2.10}$$

Under the stationarity assumption (2.9), K is the mean number of substitutions per homologous nucleotide site since divergence. Of course, if t is known, then the substitution rate can be estimated, and vice-versa.

We will now review some of the specific forms for the substitution rate matrix Q. The progenitor of these is due to Jukes and Cantor (1969).

Example 2.1

In this case, substitutions occur at the points of a Poisson process of rate λ, and when a substitution occurs it is equally likely to be to any of the other three bases. Hence

$$\underline{\pi} = (1/4,\ 1/4,\ 1/4,\ 1/4),\ Q = \lambda \begin{pmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{pmatrix}.$$

The parameter K is given by $K = 2t\lambda$.

We would like to relax the assumption of uniform base composition and equally likely substitutions.

Example 2.2

A model which retains the assumptions of uniform base composition, and a Poisson substitution scheme was proposed by Kimura (1981) to allow for different transition and transversion probabilities. The rate matrix takes the form

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} -\lambda & \gamma & \alpha & \beta \\ \gamma & -\lambda & \beta & \alpha \\ \alpha & \beta & -\lambda & \gamma \\ \beta & \alpha & \gamma & -\lambda \end{pmatrix} \end{array}$$

where $\lambda = \alpha + \beta + \gamma$. The special case $\gamma = \beta$ was studied by Kimura(1980); see also Kimura (1983, Ch. 4). Other cases in which the A and T frequencies are equal (as are the C and G frequencies) are the four parameter model of Aoki et al. (1981) and the five parameter model of Takahata and Kimura (1981).

Example 2.3

A model that allows for arbitrary base frequencies and possibly different substitution rates was proposed by Kimura (1981). This six parameter process has the form

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} \cdot & \alpha & \alpha & \alpha_1 \\ \beta & \cdot & \alpha_2 & \beta \\ \beta & \beta_2 & \cdot & \beta \\ \beta_1 & \alpha & \alpha & \cdot \end{pmatrix} \end{array},$$

the diagonal elements being determined by (2.8). Further properties of this model may be found in Gojobori et al. (1982).

Example 2.4

Felsenstein (1981), in a study of maximum likelihood methods for evolutionary trees, uses a generalization of the Jukes-Cantor model that also allows for arbitrary base frequencies. In generator form, we take $\underline{\pi}$ arbitrary, and set

$$Q = \mu \begin{pmatrix} \cdot & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \cdot & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \cdot & \pi_4 \\ \pi_1 & \pi_2 & \pi_3 & \cdot \end{pmatrix}.$$

This model corresponds to embedding a sequence of independent and identically distributed substitutions in a Poisson process of rate μ. A similar example was studied by Tajima and Nei (1982).

It has been noted by several authors (Neyman (1971), Kaplan and Langley (1979), Felsenstein (1981) among others) that the assumption of reversibility of the substitution process affords a useful simplification. Intuitively, the observation that $X(\cdot)$ (say) is reversible means that the substitution process viewed from now into the future is probabilistically identical to its behavior from now back into the past. Mathematically, the stationary Markov process $X(\cdot)$ is reversible if and only if there exists a collection of positive numbers π_j summing to unity that satisfy the balance equations

$$\pi_i q_{ij} = \pi_j q_{ji}, \quad 1 \le i, j \le 4. \tag{2.11}$$

When such exist, then $\underline{\pi}$ is the stationary distribution of the process, i.e., (2.9) holds. Reversibility is discussed at length by Kelly (1979) and Keilson (1980), for example. From (2.7) and (2.11) it follows that $\pi_i p_{ij}(t) = \pi_j p_{ji}(t)$, $1 \le i, j \le 4$, $t \ge 0$, and hence we have

$$f_{ij}(t) = \pi_i p_{ij}(2t) \tag{2.12a}$$

or, in matrix notation (cf. (2.6))

$$F_t = F_0 P_{2t}. \tag{2.12b}$$

The reversibility property is shared by several of the previous examples; it is readily checked that the process with Q matrices given in Examples 1, 2 and 4 are reversible.

This suggests that a general model incorporating the reversibility property should be studied:

Example 2.5

The generator Q of a reversible process with stationary probabilities $\underline{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4)$ can be expressed as a nine parameter matrix

$$Q = \begin{pmatrix} \cdot & x_1 & x_2 & x_3 \\ \pi_1 x_1/\pi_2 & \cdot & x_4 & x_5 \\ \pi_1 x_2/\pi_3 & \pi_2 x_4/\pi_3 & \cdot & x_6 \\ \pi_1 x_3/\pi_4 & \pi_2 x_5/\pi_4 & \pi_3 x_6/\pi_4 & \cdot \end{pmatrix} \tag{2.13}$$

satisfying $x_i \ge 0$, $1 \le i \le 6$, and the diagonal elements once more determined by (2.8).

III. ESTIMATION OF SUBSTITUTION RATES

3.1 Statistical Methods

Having described the process by which a particular homologous site evolves, we now model the stochastic structure of $\{(X_i(t), Y_i(t)), t \ge 0; i = 1, \ldots, n\}$. The simplest assumption here is that each pair of homologous nucleotides behaves independently and identically. That is, assume

$(X_i(t), Y_i(t))$, $i = 1 \ldots, n$ are i.i.d. random vectors with common distribution that of $(X(t), Y(t))$. (3.1)

It is well known that, particularly in coding regions, the structure of the base sequence is not that of independent identically distributed bases, (cf. Smith et al (1983)). As a consequence, it is customary to analyze the substitution process in coding regions according to base position in the codon. We therefore analyze three separate sequences, the first base position sequence then being $(X_{3i-2}(t), Y_{3i-2}(t))$, $i = 1, \ldots, m$, where $n = 3m$. Another reason for studying the sequences by base position in the codon involves degeneracy in the genetic code. Many substitutions in the third position of codons are silent (that is, do not change the amino-acid the codon represents). One might therefore expect heterogeneity of the substitution process along a coding region, in violation of the i.i.d. assumption (3.1).

For convenience we will denote either the whole sequences or the sequences of 1^{st}, 2^{nd} or 3^{rd} codon positions by $(X_i(t), Y_i(t))$, $i = 1, \ldots, n$. Our data now comprise the observations $N = \{N_{ij}, 1 \le i, j \le 4\}$ where

$$N_{ij} = \text{number of times we observe } X_\ell(t) = i,\ Y_\ell(t) = j,\ 1 \le \ell \le n. \qquad (3.2)$$

Under the assumption (3.1), the $\{N_{ij}\}$ have a joint multinomial distribution with parameters n, and F_t given by (2.6). Using observations on $\{N_{ij}\}$ we want to find an estimator $\hat{K}$ of the substitution parameter

$$K = 2t \sum_{\ell=1}^{4} \pi_\ell q_\ell,$$

and an estimate of the asymptotic (as $n \to \infty$) variance of $\hat{K}$.

The estimation theory for the Felsenstein model of Example 2.4 is readily elucidated. Here we have

$$K = 2t\mu H, \text{ where } H = \sum_{i=1}^{4} \pi_i(1 - \pi_i). \qquad (3.3)$$

Since $p_{ij}(t) = e^{-\mu t}\delta_{ij} + (1 - e^{-\mu t})\pi_j$, it follows that if

$$D = \sum_{i \ne j}\sum N_{ij}$$

is the number of non-identical nucleotide sites, then D has a binomial distribution with parameters n and $p = (1 - e^{-2\mu t})H$. If we assume that $\underline{\pi}$, and therefore H, is known, then the maximum likelihood estimator of K is

$$\hat{K}_F = -H \ell n\,(1 - \frac{d}{H}),\ d = \frac{D}{n}. \qquad (3.4)$$

$\hat{K}_F$ inherits its asymptotic distribution from that of d. By the 'Delta method' (cf. Serfling (1980, p. 122)) we find that

$$\hat{K}_F \sim AN(K, \frac{H^2 p(1-p)}{n(H-p)^2}), \qquad (3.5)$$

where $AN(\mu, \sigma^2)$ denotes "asymptotically normal, with mean μ and variance σ^2."

In the special case $H = 3/4$ that obtains when $\underline{\pi} = (1/4, 1/4, 1/4, 1/4)$, this model differs only notationally from Example 2.1. From (3.4) we obtain the Jukes-Cantor (1969) estimator

$$\hat{K}_{JC} = -\frac{3}{4} \ell n(1 - \frac{4d}{3}), \quad d = \frac{D}{n}, \qquad (3.6)$$

and

$$\hat{K}_{JC} \sim AN(K, \frac{9p(1-p)}{n(3-4p)^2}). \qquad (3.7)$$

The variance term in (3.7) is due to Kimura and Ohta (1972); in practice, p is estimated from the data by d. If $\underline{\pi}$ is assumed unknown, and hence must be estimated from the data, then the

variance of $\hat{K}_P$ is no longer given by the appropriate term in (3.5). The correct term can readily be calculated numerically.

The estimation theory of Example 2.2 is given by Kimura (1981). Table 3.1 describes this model in more detail.

Table 3.1

Combinations of bases

Difference	Transition type	Transversion type	
X	T C A G	T A C G	T G C A
Y	C T G A	A T G C	G T A C
Frequency	P	Q	R

With this notation, the estimator of K is

$$\hat{K}_K = -1/4 \ \ell n[(1-2P-2Q)(1-2P-2R)(1-2Q-2R)], \tag{3.8}$$

and the asymptotic variance is given by Kimura's equation [12].

Kaplan and Risko (1982) proposed an interesting alternative approach to the estimation of substitution rates. Suppose that the Q matrix is of the form $Q = \lambda(R - I)$, where $R=(r_{ij})$ is a stochastic matrix with $r_{ii}=0$, $1 \le i \le 4$. The substitution parameter is $K = 2\lambda t$, and their estimator of K is

$$\hat{K}_{KR} = 2(1 - \sqrt{1 + \ell n(1 - d)}), \ d = \frac{D}{n} \tag{3.9}$$

and

$$\hat{K}_{KR} \sim AN(K, \frac{d}{n(1-d)(1+\ell n(1-d)}). \tag{3.10}$$

This estimator was derived by approximating the form of P_t, and it should apply to cases in which d is close to 0. See also Kaplan (1983). Estimation theory for Example 2.4 is discussed in detail by Gojobori et al. (1982).

3.2 The reversible model

The final case we consider here is for a general reversible model of Example 2.5. There are essentially two different approaches to this, depending on what is assumed about $\underline{\pi}$. If we define

$$M_{ij} = \begin{cases} N_{ii} & , j = i , \\ N_{ij} + N_{ji} & , 1 \le i < j \le 4, \end{cases}$$

then under the model of (2.12) and (2.13) the joint distribution of $\{M_{ij}\}$ is multinomial, with parameters n, and $\{\ell_{ij}\}$, say, where

$$\ell_{ij} = \begin{cases} \pi_i p_{ii}(2t) & , j = i , \\ 2\pi_i p_{ij}(2t) & , 1 \le i < j \le 4 . \end{cases} \tag{3.11}$$

If $\underline{\pi}$ is assumed unknown, and hence has to be estimated from the data, the statistical problem reduces to the estimation of the nine parameters $(tx_1, tx_2, \ldots, tx_6, \pi_1, \pi_2, \pi_3)$ of the Q-matrix in (2.13) from the multinomial data (M_{ij}), with cell probabilities (ℓ_{ij}). It can be shown that the maximum likelihood estimators of these parameters may often be found by solving the equations

$$n^{-1}\bar{N} = F_0 \exp(\bar{Q}) \tag{3.12}$$

where $\bar{N}_{ij} = (N_{ij} + N_{ji})/2$, and $\bar{Q}$ is given by a matrix of the form (2.13). Computationally, this is straightforward because such Q-matrices are diagonalizable, so that $\exp(\bar{Q})$ may be computed easily; cf. Keilson (1979), p. 33-34, for example. The joint

asymptotic distribution of the estimators then follows from standard theory.

While (3.12) provides a simple method for estimating the parameters, it is not always true that (3.12) has a solution satisfying the restrictions of (2.13) (for example, if some $M_{ij} = 0$; see Table 3.2). In these cases, estimation of the parameters is more complicated.

When $\underline{\pi}$ is assumed known, as in the models of Felsenstein (1981) and Kimura (1981), then we can approach the problem in a different way. Our basic data remain the multinomially distributed observations $\{M_{ij}\}$, and the cell probabilities $\{t_{ij}\}$ determined now by just six (compound) parameters $(tx_1, \ldots, tx_6)$. We can estimate them by using, for example, minimum chi-squared estimation or least squares estimation techniques.

In either case, we arrive at estimates of the elements of Q which have a joint asymptotic distribution that is multivariate normal, the parameters of which can be estimated from the data. Hence we can also estimate the substitution parameter K. Some examples of the method are given in the next section, and detailed discussion of these and related methods appears in Tavaré and Janzen (1985).

3.3 Some data

In this section we will illustrate the results of these methods with two data sets. The genes are α-fetoprotein (Human [Morinaga et al. (1983)], Rat [Jagodzinski et al.(1981)] and Mouse [Law et al. (1981)] and Serum Albumin (Human [Dugaiczyk et al. (1982), Lawn et al. (1981)], Rat [Sargent et al. (1981)], and Mouse). Estimates of K using different estimators are given in Tables 3.2 and 3.3.

The results in Tables 3.2 and 3.3 are qualitatively similar to those found by other authors. Because of the chemical structure

of DNA and the degeneracy in the genetic code, one would expect that in coding regions the second base should have the lowest rate of acceptable substitutions, and the third base the highest rate. All the estimators give similar results either when the divergence time or when the substitution rate is small. The most noticeable differences occur for estimates of high rates, where the models with fewer parameters give lower values of K.

Table 3.2

Estimates of K (and standard deviation) for Serum albumin.*

Estimator K	Base position in codon (n = 608) 1	2	3
JC (3.6), (3.7)	.1752 (.0186)	.1387 (.0162)	.6566 (.0483)
F† (3.4), (3.5)	.1756 (.0187)	.1392 (.0163)	.6573 (.0484)
K (3.8)	.1760 (.0188)	.1389 (.0163)	.7230 (.0642)
KR (3.9), (3.10)	.1778 (.0192)	.1403 (.0166)	.6967 (.0549)
Reversible (3.12)	.1794 (.0196)	.1415 (.0169)	.7274 (.0659)

** (in the 2 column)

* Base length 1824 bases. Estimates based on Rat-Man data.

† Standard deviation assuming $\underline{\pi}$ is unknown is same as that given to 4 d.p.

**Estimation of parameters not possible by method of (3.12), since no GT or TG sites were found in data. Figures given here correspond to sequence with one GT-site added to the sequence.

Table 3.3

Estimates of K (and standard errors) for alpha-fetoprotein.*

Estimator	Base position in codon (n = 586)		
K	1	2	3
JC (3.6), (3.7)	.2298 (.0224)	.1614 (.0200)	.4840 (.0377)
F (3.4), (3.5)	.2303 (.0225)	.1921 (.0201)	.4846 (.0378)
K (3.8)	.2324 (.0229)	.1936 (.0205)	.5175 (.0452)
KR (3.9), (3.10)	.2342 (.0232)	.1945 (.0206)	.5048 (.0411)
Reversible (3.12)	.2343 (.0234)	.1967 (.0212)	.5205 (.0458)

*Base length 1758 bases. Estimates based on Mouse-Man data.
†Standard deviation assuming $\underline{\pi}$ is unknown is same as that given to 4 d.p.

IV. A STATISTICAL LOOK AT THE SUBSTITUTION PROCESS

The previous sections of this article have described in some detail a class of Markovian stochastic models that have been used to estimate substitution rates from sequence data. In this section, I want to look briefly at some statistical problems associated with the selection of classes of processes that adequately describe (in a statistical sense) the observations.

The data used in studies of the type described here involve observations taken at a single time point, t. On the basis of such data, we want to assess something of the nature of the substitution process over time. One particular question is whether the substitution process has proceeded at the same rate in both species. The following examples illustrate the possibilities.

Example 4.1

We suppose that the substitution process is described by (2.3), where $P_{Xt} = (p^X_{ij}(t))$ and $P_{Yt} = (p^Y_{ij}(t))$ are the transition matrices of Markov processes of the form (2.7). We will look at a variation of the Felsenstein model of Example 2.4, in which the generators Q_X and Q_Y are given by

$$Q_X = \mu_X Q,\ Q_Y = \mu_Y Q,\ Q = \begin{pmatrix} \cdot & \pi_2 & \pi_3 & \pi_4 \\ \pi_1 & \cdot & \pi_3 & \pi_4 \\ \pi_1 & \pi_2 & \cdot & \pi_4 \\ \pi_1 & \pi_3 & \pi_3 & \cdot \end{pmatrix}$$

where $Q\underline{1} = \underline{0}$, and $\underline{\pi}^T \underline{1} = 1$. From (2.3), we have

$$\begin{aligned} F_t &= P_{Xt}^T F_0 P_{Yt} && (F_0 = \mathrm{diag}(\pi_1, \ldots, \pi_4)), \\ &= F_0 e^{Q_X t} e^{Q_Y t} && \text{(by reversibility)}, \\ &= F_0 e^{(Q_X+Q_Y)t} && \text{(since } Q_X, Q_Y \text{ commute)}, \\ &= F_0 e^{(\mu_X+\mu_Y)Qt}. \end{aligned}$$

For this model, the mean number K of substitutions per site is

$$K = (\mu_X + \mu_Y)Ht,\ H = \sum \pi_i(1 - \pi_i).$$

An estimator of K is the Felsenstein estimator $\hat{K}_F$ described by(3.4); note that it is based solely on the number of sites showing non-identical bases. The parameters μ_X and μ_Y are confounded in the definition of K, and identical estimates of K

can arise from models with equal substitution rates or with widely different rates.

Example 4.2

A simple modification of the Jukes-Cantor process described in Example 2.1 is to make the substitutions in one gene occur at the points of a non-homogeneous Poisson process with intensity function $\lambda(u)$, $u \geq 0$, while those in the other gene occur at the points of a Poisson process of rate λ. The mean and variance of the number of substitutions per homologous nucleotide site is then $K = \lambda t + \int_0^t \lambda(u)du$. If $\int_0^t \lambda(u)du = \lambda t$, (for example, if $\lambda(u) = \lambda/2$ $(0 \leq u \leq t/2)$; $3\lambda/2$ $(t/2 < u \leq t)$) then the data will be statistically indistinguishable from those produced by the standard Jukes-Cantor process.

These two elementary examples suggest that care should be taken in making inferences about the substitution process on the basis of data taken at a single time point. However, some assumptions of the models of Sections 2 and 3 can be checked by a non-parametric approach.

To describe these methods, we return to the basic description of (2.3). Dropping the t's for notational convenience, we have

$$f_{ij} = \sum_{\ell} \pi_{\ell} \, p^X_{\ell i} p^Y_{\ell j}. \qquad (4.1)$$

Under (4.1), the marginal distribution of X is

$$f_{i+} = P[X = i] = \sum_{\ell} \pi_{\ell} p^X_{\ell i} \quad ; \; 1 \leq i \leq 4, \qquad (4.2)$$

while that of Y is

$$f_{+j} = P[Y = j] = \sum_{\ell} \pi_{\ell} p^Y_{\ell j} \quad ; \; 1 \leq j \leq 4. \qquad (4.3)$$

If, as in (2.5), $p^X_{ij} = p^Y_{ij} \equiv p_{ij}$ for all i and j then $F = (f_{ij})$ is symmetric, and the marginals of X and Y will be identical. Notice

that we only require equality of P^X_{ij} and P^Y_{ij} (for all i, j) for the single (special) time point t; recall Example 4.2.

4.1 Contingency table methods

Under the assumption (3.1) of independent and identically distributed nucleotide sites, the observation matrix $N=(N_{ij})$ defined by (3.2) has the form of a contingency table, with underlying cell probabilities $F = (f_{ij})$. Some questions of interest to our modelling problem may now be re-expressed as hypothesis tests about the structure of (two-way) contingency tables. Perhaps the most useful test of this type involves the test for symmetry of F. Several such tests have been proposed, but the simplest one for our purposes is that devised by Bowker (1948). Under the null hypothesis that F is symmetric with $f_{ij} + f_{ji} > 0$, he established that the statistic

$$X^2 = \sum_{i<j} \sum \frac{(N_{ij} - N_{ji})^2}{N_{ij} + N_{ji}} \qquad (4.4)$$

is asymptotically as $n \to \infty$ distributed as χ^2 with 6 degrees of freedom. In Table 4.1, we give observed values of the χ^2 statistic for the data used in Table 3.2 and 3.3.

Table 4.1

Observed χ^2 values for test of symmetry (4.4).*

Sequences	Base position 1	2	3
Albumin (Rat-Man)	20.17	5.49**	55.33
α-fetoprotein (Mouse-Man)	4.35	1.72	45.82

*5% significance point of χ^2_6 = 12.59

1% significance point of χ^2_6 = 16.81

**5 degrees of freedom

The results of this screening suggest that for the third position data, the Markovian models described in Section 3 are not appropriate.

Of course, we may have marginal homogeneity (that is, $f_{i+} = f_{+i}$, $\forall$ i) without symmetry. Such behavior is exhibited, for example by Markovian models in which the initial distribution of X and Y is the stationary distribution of both Q_X and Q_Y. Once more, several methods to judge the hypothesis of marginal homogeneity have been proposed. Maximum likelihood and minimum discrimination information approaches use iterative methods (cf. Ireland et al. (1969), Madansky (1963). See also the approach of Grizzle et al. (1969)). However, a simple test statistic has been described by Stuart (1955). Define

$$N_{i+} = \sum_j N_{ij},\ N_{+j} = \sum_i N_{ij} \text{ and } d_i = N_{i+} - N_{+i},\ i=1, 2, 3.$$

Let $V=(V_{ij})$ be a 3×3 matrix with elements

$$V_{ii} = N_{i+} + N_{+i} - 2N_{ii},\ V_{ij} = -(N_{ij} + N_{ji}),\ i \neq j.$$

If $\underset{\sim}{d} = (d_1, d_2, d_3)$, then the statistic

$$S^2 = \underset{\sim}{d}^T V^{-1} \underset{\sim}{d} \tag{4.5}$$

has asymptotically a χ^2 distribution with 3 degrees of freedom under the null hypothesis of marginal homogeneity. Table 4.2 gives the observed S^2 values for our data sets.

Table 4.2

Observed S^2 values for test of marginal homogeneity (4.5).*

Sequences	Base position 1	2	3
Albumin (Rat-Man)	13.19	5.19	54.86
α-fetoprotein (Mouse-Man)	2.31	1.03	45.04

*5% significance point of χ^2_3 = 7.82

1% significance point of χ^2 = 11.34

Note that the third position data exhibits high marginal inhomogeneity, suggesting once more that the Markov models analyzed in Sections 2 and 3 are not appropriate.

Note that a process in which F has neither the marginal homogeneity nor the symmetry property could still be generated by a time-homogeneous Markovian scheme, but the generators Q_X and Q_Y should be different, and $\underset{\sim}{\pi}$ cannot then be the stationary distribution for both X and Y (for then, marginal homogeneity would obtain). It is worth noting that in this case, the quantity

K in (2.10) that we have tried to estimate is no longer the mean number of substitutions per site in time t, but should be interpreted in an asymptotic sense.

Finally, even if the assumption of marginal homogeneity is reasonable, we can further test for the form of the resultant marginal distribution. For example, one property of the Kimura model of Example 2.2 is that the marginal distribution is $\underset{\sim}{\pi}$ = (1/4, 1/4, 1/4, 1/4). We may test such an assumption within our contingency table framework by testing F for given marginals (in this case, both being $\underset{\sim}{\pi}$ given above). Methods for testing for given marginals are described by Ireland and Kullback (1968), for example. These methods are, once more, iterative in spirit; rather than record the details, we give in Table 4.3 the values of the goodness-of-fit statistic for the data of first and second codon positions. The third position data are omitted, since marginal homogeneity is ruled out by the results of Table 4.2.

The results in Table 4.3 show that the data are incompatible with $\underset{\sim}{\pi}$ = (1/4, 1/4, 1/4, 1/4). From the point of view of estimating K within the Markovian framework, this might not seem to matter; in both first and second base positions, the data in Tables 3.2 and 3.3 have similar estimates of K for many underlying models. However, one question of interest involves estimation of transversion and transition rates. These estimates are based on a more detailed examination of estimates of the elements of Q, and such estimates are particularly sensitive to departures from the underlying form of $\underset{\sim}{\pi}$.

Table 4.3

Observed value of test of given marginals

$\underset{\sim}{\pi}$ = (1/4, 1/4, 1/4, 1/4) in both species.*

Sequences	Base position 1	2
Albumin (Rat-Man)	53.54	75.06
α-fetoprotein (Mouse-Man)	24.60	61.05

*5% signficance point of χ^2_6 = 12.59

1% signficance point of χ^2 = 16.81

4.2 Independence

The contingency table analyses presented here depend a good deal on the assumption that homologous sites behave independently and identically. This assumption then allows us to use standard asymptotic results for contingency tables. Several authors have studied the effects of serial dependence on the asymptotic behavior of such 'standard' contingency table test statistics. The results of these studies suggest that departures from independence can cause serious distortions in the 'usual' χ^2 tests (cf. Tavaré and Altham (1983), Tavaré (1984), Gleser and Moore (1983, 1984)). We therefore analyzed the base composition of each gene in each species of the basic data set described in Section 3.3 using standard Markov chain methods; cf. Chatfield (1973).

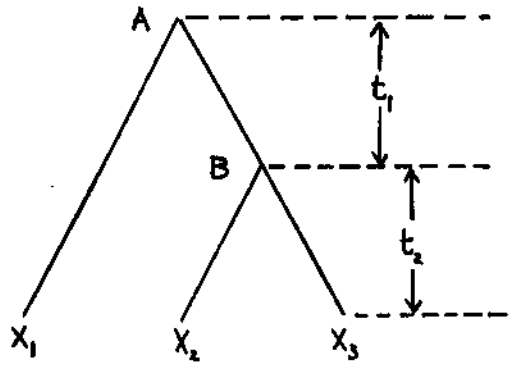
The results indicated, perhaps surprisingly, that each of the three codon position sequences is not inconsistent with independence. While marginal independence of this type is not sufficient to establish the stronger independence required in

(3.1), the results indicate that for the sequences examined here, (3.1) may not be unreasonable.

V. ESTIMATING DIVERGENCE TIME FROM A CALIBRATED TREE

Suppose now that we have homologous sequences from three species, and that the species have the known phylogeny displayed in Figure 5.1.

Figure 5.1



In Figure 5.1, X_i denotes the nucleotide appearing at a particular homologous site in species i, i = 1, 2, 3. We will assume that the phylogeny is calibrated (perhaps from the fossil record, as in Jacobs and Pilbeam (1980)), in that $t = t_1 + t_2$ is assumed known. The problem is to estimate the divergence time t_2 of species 2 and 3, and also to estimate the variance of this estimate.

For simplicity, we will assume that the substitution process leading to the observations (X_1, X_2, X_3) has the same stochastic structure in each arm of the tree, and that each process is Markovian with transition matrix $P_s = (p_{ij}(s))$, as described by one of the models in Section 2. If we define

$$f_{ijk} = P(X_1 = i, X_2 = j, X_3 = k),$$

then by conditioning on the ancestral nucleotide at positions A and B in Figure 5.1, we have

$$f_{ijk} = \sum_r \pi_r \, p_{ri}(t_1 + t_2) \sum_z p_{rz}(t_1) p_{zj}(t_2) p_{zk}(t_2), \qquad (5.1)$$

where π_r is the probability of base r at node A. If we assume once more that $\underline{\pi}$ is the stationary distribution for P_t, and P_t is reversible, then Felsenstein's (1981) "Pulley Principle" reduces (5.1) to

$$f_{ijk} = \sum_z \pi_z \, p_{zi}(t_2 + 2t_1) p_{zj}(t_2) p_{zk}(t_2). \qquad (5.2)$$

To proceed further, we assume a model like Felsenstein's given in Example 2.4. The simple structure allows us to compute (5.2) easily. Under the assumption that each homologous site behaves independently and identically, the random variables N_{ijk} given by

N_{ijk} = number of times we observe $X_1 = i, X_2 = j, X_3 = k$ in the n homologous sites

have a joint multinomial distribution with parameters n and f_{ijk}, $i \le i, j, k \le 4$. Once more using '+' to denote summation over that index, define

$$N_{12} = \sum_i N_{ii+} = \#(X_1 = X_2),$$

$$N_{13} = \sum_i N_{i+i} = \#(X_1 = X_3),$$

$$N_{23} = \sum_i N_{+ii} = \#(X_2 = X_3),$$

and set $d_{ij} = N_{ij}/n$. In the notation of (3.3), we have

$$E(\tfrac{1}{2}(d_{12} + d_{13})) = 1 - H + He^{-2\mu t},$$
$$E(d_{23}) = 1 - H + He^{-2\mu t_2}. \qquad (5.3)$$

Hence we may use as an estimator of t_2 the quantity

$$\hat{t}_2 = t \, \ell n \, \eta_2 / \ell n \, \eta_1 \qquad (5.4)$$

where $\eta_1 = \frac{1}{H}[1/2(d_{12} + d_{13}) - (1 - H)]$, $\eta_2 = \frac{1}{H}[d_{23} - (1 - H)]$, and $H = \sum \pi_i(1 - \pi_i)$ is assumed known.

The asymptotic variance of t_2 can be estimated from the asymptotic joint normality of (d_{12}, d_{13}, d_{23}) using the multivariate delta method yet again; numerical values are readily evaluated on a computer.

To give a flavor of the results, we present in Tables 5.1 and 5.2 the estimates of the divergence time of rat (X_2) and mouse (X_3) based on a tree calibrated by the known divergence time t of man (X_1) and rat, using the data for α-fetoprotein and serum albumin described in Section 3.3.

The discussion of Section 4 suggests that the assumptions made in arriving at the estimates in Table 5.1 and 5.2 may not be appropriate for the 1st and 3rd codon position data; recall the inherent asymmetry involved in 3rd positions. The second position leads to estimates of 14.6 (serum albumin) and 33.1 (α-fetoprotein) million years (MY) for the divergence time of rat and mouse. A common estimate, based on both genes, is then about 23.8 ± 4.87 MY. This figure is somewhat larger than the 8-14 MY figure suggested by Jacobs and Pilbeam on the basis of fossil evidence.

Table 5.1

Divergence time ($\hat{t}_2$) in millions of years (MY) of mouse and rat based on data from Serum Albumin gene.* Time of divergence of man and rat taken to be t = 80 MY

Base position in codon	$\hat{t}_2 \pm$ std. error
1	33.8 ± 5.91
2	14.6 ± 4.27
3	30.8 ± 3.93

*Based on 1254 homologous sites

Table 5.2

Divergence time ($\hat{t}_2$) in millions of years (MY) of mouse and rat based on data from α-fetoprotein gene. * Time of divergence of man and rat taken to be t = 80 MY

Base position in codon	$\hat{t}_2 \pm$ std. error
1	33.5 ± 5.15
2	33.1 ± 5.31
3	35.3 ± 3.93

*Based on 1551 homologous sites

There are many directions in which analyses of this sort can be extended. Most obviously, we could use other Markovian models of the type described in Sections 2 and 3. The general reversible models seem particularly tractable; a crude estimate of t_2 for the above data is 14.4 (serum albumin) and 32.6 (α-fetoprotein) MY, with an average of 23.5 MY; this differs little from the simpler Felsenstein model's results. Kaplan and Risko (1982) extend their method for two species (cf. 3.9) and 3.10)) to the case of m species with known phylogeny; their approach could easily be modified to attack the present problem, too.

From a statistical point of view, the methods described in Section 4 will apply equally well in this setting; the analyses of contingency tables suggested there carry over to three- and higher dimensional tables also. The stochastic methods here can also be extended to cover the case of 4 or more species with known phylogeny.

VI. CONCLUSIONS

This paper has given a rather bald account of the mathematical and statistical aspects of one problem in the theory of molecular evolution. Without a doubt, the mathematical models studied here are grossly simplified. Nevertheless, the vast amounts of data available on DNA sequences suggest that useful models can be developed. The statistical approaches outlined here should be useful in finding parsimonious descriptions of the data.

I have not touched on some related aspects of the central problem. In particular, there are several studies focussing on the estimation of transition and transversion probabilities; cf. Fitch (1980) and Holmquist (1983). Estimation in the Markov models discussed here provides another statistical approach that may prove useful.

One area which we are studying involves the fitting of more general models that allow for the observed asymmetry in the data of third codon positions. Such models will also allow us to assess the stability of estimates of divergence times based on sequence data; Tavaré and Janzen (1985).

The difficult and challenging problems of statistical estimation of the phylogeny itself have not been described here. Felsenstein (1983) provides an excellent overview of this area.

ACKNOWLEDGEMENTS

I would like to thank Ian Saunders for some helpful suggestions. Wen-Hsiung Li was kind enough to provide the sequence data used to illustrate the results. I also thank Hari Iyer for his discussions about the computational aspects of these problems.

REFERENCES

1. Aoki, K., Tateno, Y., and Takahata, N. (1981) Estimating evolutionary distance from restriction maps of mt DNA with arbitrary G + C content. J. Mol. Evol., 18: 1-8.

2. Bowker, A.H. (1948) A test for symmetry in contingency tables. J. Amer. Statist. Soc., 43: 572-574.

3. Chakraborty, R. (1977) Estimation of time of divergence from phylogenetic studies. Can. J. Genet. Cytol., 19: 217-223.

4. Chatfield, C. (1973) Statistical inference regarding Markov chain models. Applied Statistics, 22: 7-21.

5. Dugaiczyk, A., Law, S.M., and Dennison, O.E. (1982) Nucleotide sequence and encoded amino acids of human serum albumin in RNA. Proc. Nat. Acad. Sci., 79: 71-75.

6. Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol., 17: 368-376.

7. Felsenstein, J. (1983) Statistical inference of phylogenies. J. Roy. Statist. Soc. A., 146: 246-272.

8. Fitch, W.M. (1980) Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes. Comparison of several methods and three Beta hemoglobin mRNA's. J. Mol. Evol., 16: 153-209.

9. Gleser, L.J. and Moore, D.S. (1983) The effect of dependence on chi-squared and empiric distribution tests of fit. Ann. Stat., 11: 1100-1108.

10. Gleser, L.J. and Moore, D.S. (1984) The effect of positive dependence on chi-squared tests for categorical data. Technical Report, Dept. of Statistics, Purdue University.

11. Gojobori, T., Ishii, K., and Nei, N. (1982) Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotide. J. Mol. Evol., 18: 414-423.

12. Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969) Analysis of categorical data by linear models. Biometrics, 25: 489-504.

13. Holmquist, R. (1983) Transitions and transversions in evolutionary descent: an approach to understanding. J. Mol. Evol., 19: 134-144.

14. Ireland, C.T., Ku, H.H., and Kullback, S. (1969) Symmetry and marginal homogeneity of an r × r contingency table. J. Amer. Statist. Soc., 64: 1323-1341.

15. Ireland, C.T. and Kullback, S. (1968) Contingency tables with given marginals. Biometrika, 55: 179-188.

16. Jacobs, L.L. and Pilbeam, D. (1980) Of mice and men: fossil based divergence dates and moelcular "clocks". J. Hum. Evol., 9: 551-555.

17. Jagodzinski, L.L., Sargent, T.D., Yang, M., Glackin, C., and Bonner, J. (1981) Sequence homology between RNA's encoding rat α-fetoprotein and rat serum albumin. Proc. Nat. Acad. Soc., 78: 3521-3525.

18. Jukes, T.H. and Cantor, C.H. (1969) Evolution of protein molecules. In "Mammalian Protein Metabolism" H.N. Munro (Ed.), Academic Press, New York, pp. 21-123.

19. Kaplan, N. (1983) Statistical analysis of restriction enzyme map data and nucleotide sequence data. In "Statistical Analysis of DNA Sequence Data", B.S. Weir (Ed.), M. Dekker, New York, pp. 75-107.

20. Kaplan, N. and Langley, C. (1979) A new estimate of sequence divergence of mt DNA using restriction endonuclease mappings. J. Mol. Evol., 13: 295-304.

21. Kaplan, N. and Risko, K. (1982) A method for estimating rates of nucleotide substitution using DNA sequence data. Theor. Popn. Biol., 21: 318-328.

22. Karlin, S. and Taylor, H.M. (1975) "A First Course in Stochastic Processes." 2nd Edn. Academic Press, New York.

23. Keilson, J. (1979) "Markov Chain Models. Rarity and Exponentiality." Applied Math. Sciences, Vol. 28, Springer-Verlag, New York.

24. Kelly, F.P. (1979) "Reversibility and Stochastic Networks." J. Wiley, New York.

25. Kimura, M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol., 16: 111-120.

26. Kimura, M. (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Nat. Acad. Sci., 78: 454-458.

27. Kimura, M. (1983) "The Neutral Theory of Molecular Evolution." Cambridge University Press, New York.

28. Kimura, M. and Ohta, T. (1972) On the stochastic model for estimation of mutational distance between homologous proteins. J. Mol. Evol., 2: 87-90.

29. Law, S.W. and Dugaiczyk, A. (1981) Homology between the primary structure of α-fetoprotein, deduced from a complete cDNA sequence, and serum albumin. Nature, 291: 201-205.

30. Lawn, R.M., Adelman, J., Bock, S.C., Franke, A.E., Houck, C.M. Najarian, R.C., Seeburg, P.H., and Wion, K.L. (1981) The sequence of human serum albumin cDNA and its expression in E. Coli. Nuc. Acids Res., 9: 6103-6114.

31. Madansky, A. (1963) Tests of homogeneity for correlated samples. J. Amer. Statist. Soc., 58: 97-119.

32. Morinaga, T., Sakai, M., Wegmann, T.G., and Tamaoki, T. (1983) Primary structures of human α-fetoprotein and its mRNA. Proc. Nat. Acad. Sci., 80: 4604-4608.

33. Neyman, J. (1971) Molecular studies of evoluation: a source of novel statistical problems. In "Statistical Decision Theory and Related Topics," S.S. Gupta, J. Yackel (Eds). Academic Press, New York, pp. 1-27.

34. Sargent, T.D., Yang, M., and Bonner, J. (1981) Nucleotide sequence of cloned rat serum albumin messenger RNA. Proc. Nat. Acad. Sci. 78: 243-246.

35. Serfling, R.J. (1981) "Approximation Theorems of Mathematical Statistics." J. Wiley, New York.

36. Smith, T.F., Waterman, M.S., and Sadler, J.R. (1983) Statistical characterisation of nucleic acid sequence functional domains. Nuc. Acids Res., 11: 2205-2220.

37. Stuart, A. (1955) A test for homogeneity of the marginal distributions in a two-way classification. Biometrika, 42: 412-416.

38. Tajima, F. and Nei, M. (1982) Biases of estimates of DNA divergence obtained by the restriction enzyme technique. J. Mol. Evol., 18: 115-120.

39. Takahata, N. and Kimura, M. (1981) A model of evolutionary base substitutions, and its application with special reference to rapid change of pseudogenes. Genetics, 98: 641-657.

40. Tavaré, S. (1983) Serial dependence in contingency tables. J. Roy. Statist. Soc. B., 45: 100-106.

41. Tavaré, S. and Altham, P.M.E. (1983) Dependence in goodness of fit and contingency tables. Biometrika, 70: 139-144.

42. Tavaré, S. and Janzen, T. (1985) On estimating substitution rates from pairs of nucleotide sequences. Mol. Biol. Evol., in preparation.

DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
FORT COLLINS, COLORADO